DOCUMENT RESUME

ED 190 624 TM 800 418

AUTHOR	Taylor, Hugh
TITLE	Grading Practices: Issues and Alternatives.
INSTITUTION	Victoria Univ. (British Columbia).
SPONS AGENCY	British Columbia Dept. of Education, Victoria.
PUB DATE	Aug 79
NOTE	50p.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; Data Analysis; *Evaluation Criteria; Foreign Countries; *Grades (Scholastic); *Grading; *Guidelines; Reliability; *Scores; Student Evaluation; Weighted Scores
IDENTIFIERS	Canada

ABSTRACT

Various practices of grading, and some of their advantages and disadvantages are discussed. The first chapter presents the purposes and meaning of letter grades, and offers some guidelines: letter grades should reflect academic achievement, rather than affective behavior or personality: they should be closely related to course objectives and learning outcomes, and should be determined by a wide range of appropriate information sources. The second chapter describes several procedures for assigning letter grades, based upon: (1) percentage scores predetermined before testing: (2) group performance--normal curve using percentiles or standard deviations: (3) natural breaks, or gaps between scores: (4) group ability (score frequency): (5) growth or improvement: (6) performance contracts: and (7) criterion-referenced grading. Suggestions for combining evidence of achievement involve assigning relative weights to different data components, or combining components, using standard deviations or stanines. Ways to insure comparability of grades across teachers and subject areas include publication and discussion of letter grade distributions, use of an anchor test, and graphing of all assigned grades. Twenty-three discussion questions are included, as well as instructions for converting a set of scores into stanines, a 28-item glossary, and an 11-item bibliography. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190624

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Grading Practices: Issues and Alternatives

TM 800418

Province of
British Columbia
Ministry of Education,
Science and Technology
LEARNING ASSESSMENT BRANCH

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
P. Northover

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

GRADING PRACTICES:

ISSUES AND ALTERNATIVES

Prepared by

Hugh Taylor
University of Victoria

For the

Learning Assessment Branch
Ministry of Education, Science and Technology
Province of British Columbia

August 1979

JUN 2 0 1980

GRADING PRACTICES: ISSUES AND ALTERNATIVES

Errata

Page 17 - Figure 1 - between 69%ile and 93%ile - and above 24%: read Letter Grade B for Letter Grade P.
- between 93%ile and positive end of curve and above 7% read Letter Grade A for Letter Grade F.

Page 21 - The material on this page should be read following paragraph 1 on page 22.

Page 22 - Paragraph 1, last line: read 'from' instead of 'form'.

Page 34 - Last line: read 'beyond' instead of 'below'.

PREFACE

There is probably no topic in education which has evoked as much sustained discussion and controversy, both historically and currently, as that of "grading".

A number of approaches to and procedures for grading have resulted from these discussions over the years. One of the most familiar and most widely used approaches has as its basis a five- or six-point letter grade scale (i.e., A, B, C, P, F). Because of its wide use in British Columbia schools, this approach to grading is discussed in detail in this resource book as further information for those using letter grades.

This resource book is intended to provide a descriptive rather than a prescriptive treatment of this approach to grading. It can serve as a useful base for informed discussion through its provision of details on issues such as the purposes served by letter grades, the methods, advantages and disadvantages of various approaches for establishing letter grades as well as techniques for combining grade components and for examining grade comparability. It has the potential of provoking interest and reconsideration of grading methods in a school or system, and of stimulating new thought in the light of knowledge of current practices.

Appreciation is extended to Dr. Hugh Taylor, University of Victoria, for his thoughtful and comprehensive treatment of this difficult topic. The valuable contributions made by Mr. Les Dickason during the preparation of the final manuscript are gratefully acknowledged.

R. Nancy Greer, Director
Learning Assessment Branch
Ministry of Education,
Science and Technology.

August, 1979

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES 6

INTRODUCTION 7

Chapter Page

1. THE LETTER GRADE

1.1 What is a Letter Grade? 9
1.2 Purposes of Letter Grades 10
1.3 Meaning of Letter Grades -Some General Principles 11

2. BASES FOR ASSIGNING LETTER GRADES

2.1 Grading Based in Predetermined Percentage Scores 15
2.2 Grading Based on Group Performance 17
(a) Method 1 Normal Curve Using Percentiles
(b) Method 2 Normal Curve Using Standard Deviations
2.3 Grading Based on Natural Breaks 25
2.4 Grading Based on Group Ability 26
2.5 Grading Based on Growth or Improvement 28
2.6 Grading Based on Contracts 30
2.7 Grading Based on Criterion-Referencing 32

3. COMBINING EVIDENCE OF ACHIEVEMENT

3.1 Assigning Relative Weight to Components 35
3.2 Combining Components 35

4. GRADE COMPARABILITY 39

5. DISCUSSION QUESTIONS 43

APPENDIX: CONVERTING A SET OF SCORES INTO STANINES 45

GLOSSARY 49

BIBLIOGRAPHY 53

LIST OF FIGURES AND TABLES

Figure — Page

1. A Normal Distribution of Letter Grades Based on a Fixed Percentage Within Each Category 17
2. A Normal Distribution of Letter Grades Based on Standard Deviation Units 19
3. Positively and Negatively Skewed Distributions 22
4. Graphical Distribution of Scores for Natural Breaks Method 25
5. Grades Assigned by Natural Breaks Method 25
6. Example of a Letter Grade Contract for a Science Course at the Junior Secondary Level 31
7. Example of Basing Letter Grades on Criterion-Referencing for Social Studies at the Junior Secondary Level 33
8. Histograms Showing Internal and External Grade Distributions 41

Table — Page

1. Frequency Distribution of Scores and Grades Based on Percentiles 18
2. Frequency Distribution of Scores and Grades Based on Standard Deviation Units 20
3. Grade Distribution Based on Two Normal Curve Methods 23
4. Grade Distribution Based on Group Ability Estimates 27
5. Using Unweighted Raw Scores to Determine a Total 35
6. Using Weighted Raw Scores to Determine a Total 36
7. Internal and External Grades for Ten Students 40

INTRODUCTION

Assigning letter grades to pupils is a major concern of teachers at all levels of education. This concern has heightened with the recent emphasis on individualized instruction, mastery learning programs, and other innovative teaching and organizational systems. However, much more basic concerns have been the practical and philosophical ambiguities with which teachers approach their grading tasks. The practical aspects of grading involve, among others, such questions as the following:

What aspects of pupil behaviour should be included in determining letter grades?

How should cut-off points between different letter grades be calculated?

How can grades be made more valid and reliable?

Should grades be comparable across different teachers, schools, and subject areas?

Further, a basic philosophical question confronts some teachers who find it morally difficult to assign grades because they find the roles of friend or counsellor and judge to be incompatible. They are strongly influenced by the sentiments of the Sermon on the Mount: "Judge not, that ye not be judged."

Within this milieu, it is not surprising that the pros and cons of grading have been debated for over eighty years. Despite periodic protestations by critics and calls for its abandonment, the grading system has been maintained by educators and is deeply embedded in the educational culture. The reason appears to be relatively simple. By any criterion, the impact of a program of studies results in a broad range of achievement among members of most school classes. School and university personnel, parents, pupils and employers request evidence to show where the individual student stands on relevant achievement criteria. Valid letter grades can provide some of the evidence.

The purpose of this booklet is not to recommend any one system of assigning grades that will satisfy all teachers. Rather, it is hoped that the booklet will provide reasonable grading alternatives so that teachers may choose procedures that are in harmony with the philosophy and goals of their school, and that will in addition, alleviate many of the current shortcomings such as grade inflation, the lack of meaning of letter grades, and the problem of controlling extremely deviant grade distributions across comparable classes. It should be stressed that no grading system can be effective unless it is combined with a clear and unambiguous reporting system. Thus, the issues in this booklet must be examined in conjunction with the reporting procedures used by teacher, school and district.

CHAPTER 1

THE LETTER GRADE

The following discussion provides an orientation to the complex field of grading. First a working definition is given for the term "letter grade", along with an explanation and amplification of the major aspects of the term. When one becomes involved deeply in the details of grading, it may be helpful to refer to the definition for perspective and guidance.

1.1 What is a Letter Grade?

For the purpose of this booklet, a letter grade is defined as:

1) a summary symbol
2) evaluating a substantial segment of achievement which is
3) attained by a pupil in a course and
4) assigned by a teacher
5) for the purpose of record and report.

The first attribute, a letter grade is " a summary symbol", serves to differentiate it from a score on a test. For our purposes, to simplify the presentation, grades will be restricted to the letters, A, B, C, P and F, knowing that in many school systems other letters are used with or without the + and - modifiers.

The second and third attributes, a letter grade "evaluates a substantial segment of achievement which is attained by a pupil in a course", suggest that a letter grade should not be assigned to individual papers, specific tests or other student tasks but should be reserved for judgements that include evidence available on student achievement in a complete course or a segment of a course such as a reporting period. As such, letter grades are usually derived from the evidence available in a set of scores for each pupil. Ideally, this set of scores is developed by a teacher after using a wide variety of techniques that measure different course goals.

The fourth attribute, a letter grade is "assigned by a teacher", implies that a teacher has a systematic procedure for assigning grades. Briefly, this procedure would define clearly those aspects of pupil performance that the grade should try to characterize, as well as to what standard and/or reference group the pupil is being compared.

The final attribute, a letter grade is "used to record and report", implies that a letter grade has the character of being a permanent record, usually retained for a long period in the school record system and used to communicate with individual students, parents, teachers, other educational authorities and potential employers. If this communication is to be effective, the reporting format must be unambiguous, and must specify the basis on which the grades were determined. Without a match between the grading system the reporting format, the information communicated may be misleading.

1.2 Purposes of Letter Grades.

As stated earlier, the basic purpose of a grade is to record and report pupils' educational achievement. This basic purpose can best be explained from the standpoint of the various users of the grade: pupils, parents, teachers, other secondary school and post-secondary personnel, as well as prospective employers.

The primary consumers of grades are pupils. Letter grades should facilitate the learning and educational development of each pupil by functioning as an incentive for the improvement of academic performance, and by identifying areas of special strength and weakness. Thus the grade forms a basis for realistic self-appraisal and for future educational and vocational planning.

Letter grades, which are issued rather infrequently during the school year, are not, of course, the only means of providing feedback to pupils on their progress in a course. Day to day evaluations such as test scores, comments on papers, pupil-teacher conferences, and ratings of projects and assignments are all important sources of information that aid pupils in adjusting their short-term learning strategies. However, the teacher's summative appraisal of the pupil's progress in the course, expressed as a letter grade, is one of the most valuable records for helping the pupil make long-term educational plans. This is particularly true if the record becomes cumulative as the pupil progresses through school and provided that a common basis for grading has been used from year to year.

Since the attitudes and opinions of parents play a vital role in shaping the school progress and educational and vocational plans of their children, it is important that the parents understand the specific meaning of letter grades in all the subject matter areas. In order to accomplish this understanding, school personnel should consider parent-education meetings which stress the interpretation of school programs and grading procedures. These meetings should better prepare parents and teachers to co-operate in promoting the maximum development of the individual student. One issue of the school bulletin or newsletter each year might stress interpretation of grades, and thereby supplement such meetings.

Letter grades provide useful data for making a variety of decisions within the education system. Among these are administrative decisions related to promotion, placement within particular tracks or streams, programs or curricula, acceleration, graduation, transfer, university and college admission, honor awards and scholarships. Counselling decisions, particularly those dealing with educational planning, rely heavily on the students' cumulative grade record as a basic source of information.

At times, prospective employers request a school report and other personal information about pupils. At such times, the school personnel have the obligation to obtain the pupils' and/or parents' consent prior to forwarding the requested information. Furthermore, when school authorities forward such infromation, they have the responsibility to ensure that employers understand the meaning of the school grading system and can interpret the report correctly.

Of course, not all purposes of grades will receive equal emphasis at the elementary, secondary or tertiary levels of education. For any particular school or district, the balance in emphasis must be determined co-operatively by the school staff within the framework of the district and the school philosophy and the school district policy on assessment.

1.3 Meaning of Letter Grades.

Bringing meaning to grades so that they may be interpreted correctly by all who use them is one of the educator's most difficult tasks. Not only are there various purposes to be served by grading practices, but there are also choices among the bases for assigning grades. The problem is yet further confounded because there is not a universal agreement among teachers at any one grade level, or even among those who plan and teach the same course, as to what evidence should be collected, combined and used for grading purposes.

The problem of bringing meaning to grades involves two main questions: what type of student behaviour should be represented by a letter grade, and what type of data should be obtained to show evidence of this behaviour? A suggested approach to the first question will be outlined in the series of general principles listed below. These principles will form a basis for suggesting answers to the second question in later sections of the booklet.

Some General Principles:

1. *Letter grades should be used as judgements of academic achievement only.*

In other words, the letter grade should inform pupils and others of how successful they have been, both quantitatively and qualitatively, in achieving the goals and learning outcomes of the course. Achievement, in this sense,

means the pupil's present status -- what he knows and can do at the time the grade is reported. As grades are a means of communicating a judgement or evaluation of some aspects of pupil behaviour, they should be as unambiguous as possible. Limiting grading procedures to reporting current achievement appears to minimize the misinterpretation of grades and greatly reduce any ambiguity associated with them.

2. Letter grades should not be used to report affective behaviour or other aspects of a pupil's personality.

Implied in the first principle is the corollary that letter grades should not be used to report personality traits and characteristics. This principle eliminates the reporting on certain, possibly important, types of student behaviour through the use of letter grades. These include such aspects of behaviour as: attendance, attitude, cleanliness, conduct, conformity, discipline problems, deportment, effort, interest, personality, punctuality, work habits and others.

A letter grade based on a kind of composite reward for ability to get along well with others, for good behaviour, for trying hard, and for actual achievement of instructional objectives is almost entirely useless. Such a combination dilutes the meaning of a grade to such an extent that its reliability approaches zero and its validity becomes uninterpretable. Teachers, of course, should take notice of these aspects of student behaviour, and, when appropriate, try to communicate the information to others. However, it is better not to compound these qualities within the letter grade system. As most of those aspects of behaviour are very difficult to measure and, at times, may be greatly influenced by the personal biases of the teacher, it would appear to be a better policy to make separate ratings of these characteristics for possible inclusion in the over-all personal and social evaluation of pupils.

3. Letter grades should bear a close relationship to the course goals and learning outcomes.

Individual teachers or groups of teachers should publish a detailed outline explaining the meaning of various letter grades in terms of pupil tasks and achievement levels. Pupils should know which factors will be weighted in grading and know their relative importance. Where possible, the various letter grades should also be related in a clear and orderly fashion to the general and specific objectives of the course, since these are the basis for both the instruction and the measurement of achievement in an effectively conducted classroom. Ideally, what each letter grade means in a specific subject area should be known prior to the beginning of a course.

4. A wide range of relevant and appropriate sources of information should be used in determining letter grades.

Depending upon the course objectives and subject matter, a teacher should, whenever possible, use a variety of measurement techniques for judging levels of achievement. Among those to consider are the conventional pencil and paper tests, check lists, rating scales, projects, oral and written reports, and other assignments. Any single source of information about pupils is likely to provide a less representative estimate of their achievement level than some combination of the techniques used to obtain achievement data. Procedures for weighting and combining different types of information will be discussed in a later section of the booklet.

CHAPTER 2

BASES FOR ASSIGNING LETTER GRADES

As stated earlier, the purpose of grading is to report information about the performance and achievement of pupils relative to the goals and learning outcomes of a course or program. Before reporting, however, a teacher must develop a frame of reference for the value judgements required. For this purpose, one can choose from a number of different grading systems. In this section of the booklet, some of the most popular systems are explained and discussed in historical order, and two of the newer methods are presented.

2.1 Grading Based on Predetermined Percentage Scores.

One of the oldest, although at present less frequently used, methods of assigning letter grades is to convert test scores, or cumulative scores based on a number of different tests and/or other assignments, into percentages and then to assign letter grades on the basis of predetermined levels such as the following:

A - 80% to 100%

B - 70% to 79%

C - 60% to 69%

P - 50% to 59%

F - below 50%

This system is used in reporting grades by at least half of Canada's largest universities. It is also popular with beginning teachers and with parents who feel the system is clear and unambiguous. However, as will be shown, as a method of assigning grades, it has many problems.

One of the difficulties encountered by teachers when they set the percentage equivalents of different grades prior to the completion of tests, is that tests have to be designed so that the results fit the teacher's pre-conceived notion of what constitutes a "good" distribution of grades. If students are told that they must obtain at least "50% in order to pass", then the teacher must try to design the test so that few, if any, students obtain a score of less than fifty percent. It should be noted that when the test results turn out to be markedly different from the teacher's expectations, nefarious techniques are used to overcome the embarrassing situation. Often, a choice is made from a number of alternatives such as adjusting the marks by applying a linear equation, administering an alternate test, making the next test easier or harder to balance the effect of the first one, or perhaps changing the grading percentage equivalents.

All of these highly unprofessional activities can be avoided if grading standards are judiciously set <u>after</u> the data are available.

Another problem encountered by users of the predetermined percentage-grade system, particularly at the secondary school level, is the lack of an adequate frame of reference for interpreting the percentage score. A "50% correct" score on a test has no inherent meaning. On a test with difficult items, 50% correct could be considered a high mark, whereas on a test of relatively easy items, it might indicate a very low level of performance. The meaning of a particular percentage, therefore, varies depending upon the difficulty of the test items. When one asks a teacher what fifty percent means in terms of subject matter, a satisfactory answer is unlikely due to the lack of a clearly defined set of learning outcomes.

Using predetermined scores to assign grades, of course, causes the minimum of mental strain since no subjective decisions are required. Teachers who use such a system can operate with a simple mechanical procedure. However, this technique assumes a precision that is unrealistic and unattainable even by the most highly qualified test specialists. To obtain precision or reliability in tests, it is recommended that the average or mean score should be approximately halfway between a perfect score and a chance score. To take a simple example, if a teacher constructs a 100 item true-false test, a student should be able to obtain a score of fifty just by guesssing. Although the possible range of scores is from zero to 100, the practical range is from 50 to 100. Ideally the mean score for a class is halfway between the perfect score (100) and a chance score (50) -- in this instance, 75. Under such circumstances, there is a greater possibility for scores to spread out and show maximum individual differences within the class -- a condition that would result in greater teacher confidence in the assignment of letter grades. Most tests and test questions developed for classroom use do not attain this level of precision.

2.2 Grading Based on Group Performance

A common method for assigning grades is to judge the achievement of students in relation to their position within a peer group. This method is usually referred to as comparative or normative grading, relative grading, or grading "on the curve". These terms have a variety of meanings, among which are the following:

2.2(a) Method 1 Normal Curve Using Percentiles. One procedure for assigning grades involves giving a predetermined fixed percentage of A's and B's, etc. corresponding to the percentiles shown in Figure 1. In this case, the percentiles divide the normal distribution into five more or less equal distances along the score base line.

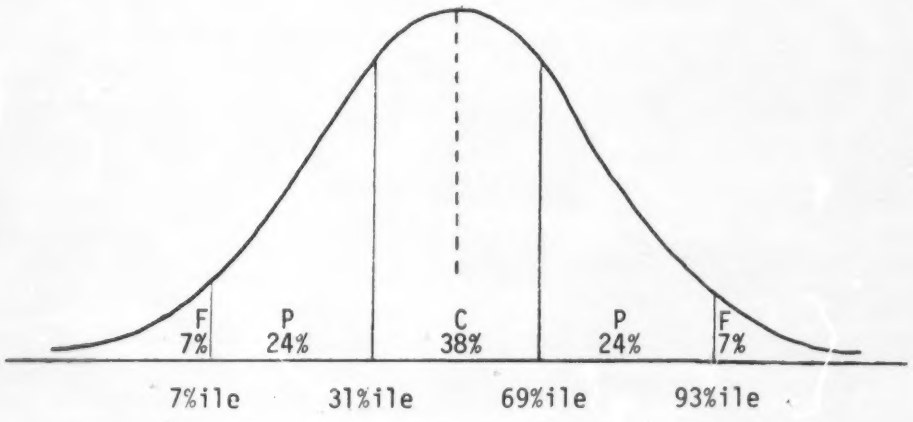


Figure 1. A Normal Distribution of Letter Grades Based on a Fixed Percentage Within Each Category.

When using this procedure, a teacher simply organizes the total scores for a class into a frequency distribution and then calculates the various cut-off scores associated with the four percentile points. For example, in Table 1, a frequency distribution for a class of 30 students is shown. Using the approximations shown in Figure 1, students scoring less than 62 (the 7%ile) would be assigned F, students scoring from 62 to 66 would be assigned P, students scoring from 67 to 73 would be assigned C, students scoring 74 to 77 would be assigned B, and students scoring 78 (the 93%ile) or higher would be assigned A. The distribution of thirty grades using this method is also shown in Table 1.

All data used in this chapter have been simplified for illustrative purposes. Such symmetrical distrubutions are unlikely to occur in real classrooms.

TABLE 1

FREQUENCY DISTRIBUTION OF SCORES AND
GRADE DISTRIBUTION BASED ON PERCENTILES

Score	Frequency	Grade Assigned
80	1	A
79	1	A
78	-	-
77	1	B
76	2	B
75	1	B
74	3	B
73	1	C
72	1	C
71	2	C
70	4	C
69	2	C
68	1	C
67	1	C
66	3	P
65	1	P
64	2	P
63	1	P
62	-	-
61	1	F
60	1	F
	30	

Grade	Frequency
A	2
B	7
C	12
P	7
F	2
	30

Of course, it is not necessary to have the original scores normally distributed nor to have a symmetrical assignment of grades. In the figure on the previous page, and in the example which follows a normal curve is used because its characteristics are well known. A non-symmetrical distribution of grades such as the following could be used if it appeared to be reasonable for a particular group: A-15%, B-20%, C-50%, P-10%, F-5%. By adjusting these figures, a teacher could eliminate the automatic predetermined percentage of failures.

2.2(b) Method 2 Normal Curve Using Standard Deviations.

An alternative to Method 1, requiring more statistical calculation, is to base letter grades on their relative distance from the mean of the scores with the unit of measurement being the raw score standard deviation (SD)*. This grading procedure involves dividing the base line of the normal curve into five equal steps along an interval scale as follows:

$$\frac{\text{5 SD Units}}{\text{5 grading categories}} = \text{1.0 SD Unit for each grade category.}$$

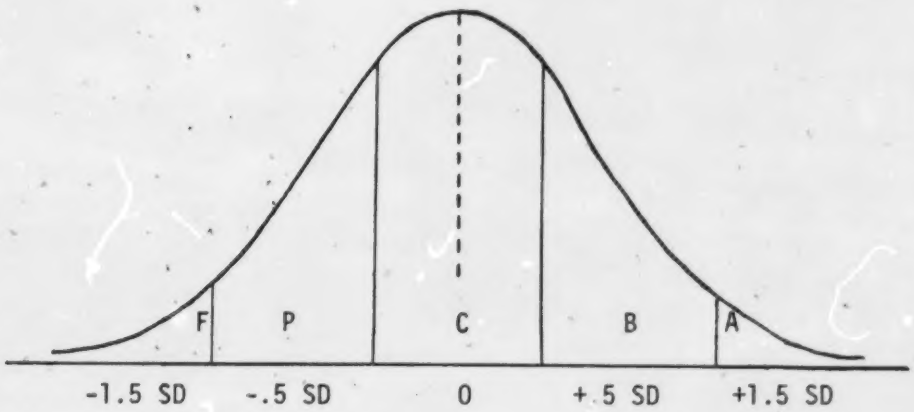


Figure 2. A Normal Distribution of Letter Grades Based on Standard Deviation Units.

*See the glossary for an explanation of terms used.

As shown in Figure 2, the method assigns A to students whose score is at or above 1.5 standard deviations above the mean (i.e. with a z score* of +1.5 or more); B to those whose scores lie between 0.5 and 1.5 standard deviations above the mean (i.e. with a z score between +0.5 and +1.5); C to those with z scores between +0.5 and -0.5; P to those with z scores between -0.5 and -1.5; and F to those with z scores at or below -1.5.

The distribution of scores for the 50 students represented in Table 2 has a mean of 40 and a standard deviation of approximately 4.4. Using this method, a score of more than 6.6 above the mean (one and one half times the standard deviation) would be required for an A (in practical terms a score of 47 or higher in this example). Rounding to whole numbers, a score between 2 and 7 above the mean would be required for a B (43 to 46); scores from 2 above the mean to 2 below the mean would be assigned a C (38 to 42); scores between 2 and 7 below the mean would be assigned P (34 to 37); and scores 7 or more below the mean would be assigned F (33 or lower). The distribution of grades using this method is also shown in the table.

TABLE 2

FREQUENCY DISTRIBUTION OF SCORES AND GRADE DISTRIBUTION BASED ON STANDARD DEVIATION UNITS

Score	Frequency	Grade Assigned
50	1	A
49	-	
48	2	A
47	1	A
46	2	B
45	2	B
44	4	B
43	3	B
42	4	C
41	4	C
40	4	C
39	4	C
38	4	C
37	3	P
36	4	P
35	2	P
34	2	F
33	1	F
32	2	F
31	-	
30	1	F
	50	

Grade	Frequency
A	4
B	11
C	20
P	11
F	4
	50

*See the glossary for an explanation of terms used.

For example, prior to instruction, it is to be expected that scores on tests would be positively skewed as shown in Figure 3a. After instruction, however, it is to be expected that scores on tests would be quite the reverse and be negatively skewed as shown in Figure 3b. Clearly these curves are substantially different from "normal". Consequently, grading on the assumption of a normal distribution of scores misrepresents reality.

a) Before Instruction: Positively Skewed

b) After Instruction: Negatively Skewed

Figure 3. Positively and Negatively Skewed Distributions.

A critique of normal curve methods for grade distribution. Each of the methods above uses the normal curve as the basic model for classifying various grades. There are, however, no rational or statistical reasons why teachers should use the normal curve approach or even assume units of equal length in their grading procedures. It is true that when a large and randomly selected group of pupils at a given grade level write a standardized test, the scores tend to form a normal distribution. Standardized test results turn out to be normally distributed as a by-product of the procedures used for selecting the test items. Having the results of standardized tests distributed normally helps a teacher interpret various derived scores (percentiles, standard scores, stanines etc.) because there is a constant relationship between these scores when they are all equated to a normal distribution. However, this does not imply that class achievement test results, or any other measured cognitive traits are normally distributed. In fact, one would hope that following instruction, a group of students would produce a set of test scores that would differ considerably form a normal distribution.

It should be noted that the use of either *Method 1* or *Method 2* would result in precisely the same assignment of letter grades, providing the score distribution is a normal one. However, if the original score distribution is irregular or skewed either positively or negatively, an individual pupil could receive a different grade depending upon the method used. Even in a very negatively skewed distribution, *Method 1* would automatically assign A's to 7% of the students. *Method 2,* however, if applied to a negatively skewed distribution may result in no A grades (if no scores were more than 1.5 standard deviation units above the mean) although, as a result there could be a preponderance of B grades.

For example, consider the negatively skewed distribution in Table 3. Calculating the grades based on the two normal curve methods results in two quite different grade distributions. Using *Method 1* seven A grades are assigned, whereas with *Method 2*, because there are no scores more than 1.5 standard deviations above the mean, no A grades are assigned. Of crucial importance to some students, *Method 2* assigns failing grades to 4 more students than *Method 1.*

In summary then, the principal difference between the methods is that *Method 2*, does not guarantee in advance that a fixed percentage of pupils will receive a given grade. As can be seen in Table 3, with *Method 2,* 7 percent of the students do not necessarily receive A nor do 7 percent automatically fail.

TABLE 3

GRADE DISTRIBUTIONS BASED ON TWO NORMAL CURVE METHODS

Score	Frequency	Grades Assigned by Normal Curve Methods: Method 1 using percentiles	Method 2 using standard deviations
60	7	A	B
59	9	B	B
58	9	B	B
57	3	B	B
56	3	B	B
55	8	C	B
54	8	C	C
53	4	C	C
52	4	C	C
51	4	C	C
50	4	C	C
49	4	C	C
48	2	C	P
47	4	P	P
46	4	P	P
45	4	P	P
44	4	P	P
43	4	P	P
42	4	P	F
41	3	F	F
40	3	F	F
39	1	F	F
	100		

Mean ≈ 52

S.D. ≈ 6

Method 1		Method 2	
Grade	Frequency	Grade	Frequency
A	7	A	0
B	24	B	39
C	38	C	28
P	24	P	22
F	7	F	11
	100		100

Another problem that confronts a teacher using the normal curve methods is how to assign grades in small homogeneous classes. It seems reasonable that honor students enrolled in advanced secondary school courses should not be limited by a method that assigns C as an average grade. Such a procedure would certainly be discriminatory because many universities admit students, at least partially, on the basis of grade point averages. Similarly, students in lower or middle ability groups should not be expected to receive a set percentage of A grades, assuming that their level of achievement is considerably lower than the high ability group. It is also difficult to divide a small homogeneous group into five grade categories. Usually there is little achievement variability within such a group, and grade categories have to be based on relatively insignificant differences in scores.

A third criticism of the two methods is that they do not take into account the over-all achievement level of the students in the class. The top test-taker in Canadian History, for example, may still not know much about the country's past, even though this student has outperformed all classmates. Using these normative methods may force a teacher to award high grades for what might be low achievement. The converse is obviously true, and students may be assigned F despite a high level of performance.

Perhaps another way of stating this criticism is to point out that when normative grading is used, a teacher is relieved of the responsibilities of having to specify acceptable levels of student achievement. A particular class may have achieved at a very high level, with all members of the class well prepared to proceed to the next instructional sequence. However, with these two normative methods, some students will be assigned low, and failing grades. In another class where achievement is generally low, some students will still receive high grades.

A further criticism of normative grading is that it sometimes creates undue competition and poor attitudes among pupils. Pupils who consistently make high scores on tests are often envied but not always liked by their peers. Normative grading sometimes causes pupils to hope that others will perform poorly on tests and assignments. This does not appear to be an appropriate attitude to cultivate.

Finally, the assumptions of normality of achievement and of symmetrical grade distributions are, at best, extremely tenuous. It is very doubtful if typical classroom test data or cumulative achievement scores are ever distributed normally -- particularly with groups smaller than 100 and, therefore, it seems inappropriate to use the methods with small class groupings.

2.3 Grading Based on Natural Breaks.

A method of assigning grades that avoids the rigidity and overcomes some of the criticism of the normal curve approach is called the "natural breaks method" or assigning grades by visual inspection. The method can be considered normative because the grades indicate how a student achieves in relation to other students. However, the procedure does not assume that grades should be based on a normal distribution, or that any fixed percentage of students should fail or receive any particular grade.

The procedure for using the natural breaks method is relatively simple. Set up on lined or graph paper a horizontal scale with equal intervals to represent the approximate range of scores. Mark off the scale to include all possible scores within the distribution. Graphically record each student score by placing an 'x' in the appropriate position. For example, in Figure 4, the range extends from 140 to 180 to accommodate the scores of the thirty-six students represented by the 'x's.

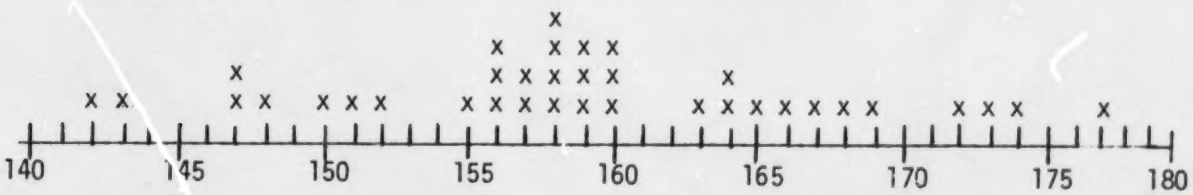


Figure 4. Graphical Distribution of Scores for Natural Breaks Method

After the scores have been set in a graphical distribution, they may be inspected to determine the grade cut-off points. There is no standard procedure for dividing the distribution. The most common approach is to let the class average be a C (or some other value) and, if possible, let the gaps or breaks in the distribution separate one grade category from another. Figure 5 shows that the middle fifty percent of the scores (from 155 to 160) was assigned a C. Other cut-off points indicated in Figure 5 were based upon the gaps in the distribution along with the best judgement the teacher could make concerning, among other factors, the general class ability, standards used in the past, as well as the type and quantity of assignments and tests used during the course.

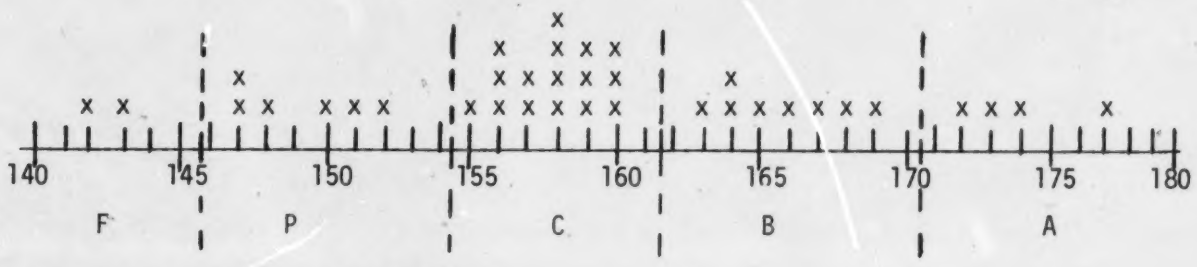


Figure 5. Grades Assigned by Natural Breaks Method

In contrast to the two previous methods that assign grades on a more or less statistical basis, the natural breaks method relies more on judicious decisions by the teacher as to what constitutes high and low quality achievement. The process is not without drawbacks, the most obvious being the dilemma facing a teacher when the distribution has no natural breaks.

2.4 Grading Based on Group Ability.

This method of assigning grades is different than those previously considered because it takes into consideration the ability levels of the pupils in a particular grade or class. In general, this method assigns a letter grade by comparing a pupil's achievement with that of other members of the group as did the previous methods, but adds the condition that the proportion of different grades assigned is predetermined by the ability levels within the group.

Such a procedure assumes that school personnel have some systematic way of estimating the general ability level of their pupils. Normally this information would be obtained either from standardized achievement test scores in a subject area, or from general ability tests, or a combination of both. Other procedures for estimating ability might include past performance data, usually letter grades, available in the student record files. When standardized test scores are used, a teacher or administrator can tabulate the percentile ranks of students in the class using the norms in the test manual to determine how the class differs from the "average".

The suggested frequency distribution of grades for a class of 30 students might be determined as follows. First, a tabulation of standardized test scores for the students in the class would be completed:

Percentile Rank Categories*	Frequency Within Class
94 and above	4
70 - 93	8
32 - 69	16
8 - 31	2
and below	0

*The percentile rank categories were chosen to correspond to those in Figure 1 on page 17.

Secondly, previous letter grade records might be explored to indicate ability levels in the class as follows:

Grade Average Based on Previous Achievement	Frequency Within Class
A	5
B	7
C	14
P	4
F	0

A teacher using these estimations which show how the talent in the class is distributed, might assign 4 or 5 A's, 7 or 8 B's, 14 to 16 C's, and 2 to 4 P's. With this scheme, one need only count down the distribution of composite scores to determine where the cut-off points between grade levels should occur.

Table 4 shows the result of using the above estimates to assign grades for the class of 30 students. The estimates should serve solely as guidelines, supplemented by the best judgements of the teacher.

TABLE 4

GRADE DISTRIBUTION BASED ON GROUP ABILITY ESTIMATES

Score	Frequency	Grade Assigned
100	1	A
99	-	-
98	2	A
97	-	-
96	1	A
95	1	B
94	-	-
93	2	B
92	2	B
91	3	B
90	2	C
89	1	C
88	2	C
87	2	C
86	4	C
85	2	C
84	-	-
83	1	C
82	-	-
81	-	-
80	1	P
79	2	P
78	-	-
77	-	-
76	-	-
75	1	F
	30	

Grade	Frequency
A	4
B	8
C	14
P	3
F	1
	30

Teachers, when using the "ability estimate" system, should not consider the grades or percentile ranking assigned to individuals during the original estimation process as expectation levels for the students so categorized. In Table 4, the four students assigned A's are not necessarily the same students who scored above the 94%ile on the standardized test. The five students categorized as "A" based on previous achievement should not be expected to achieve at the "A" level. There are many reasons why one should not attempt to predict specific grade expectations for individual students. Among them are the following:

1) There is no rational reason why, in using the foregoing procedures to determine grade distributions for a group, one should assume that the classification for an individual suggests capacity or "expectation". One cannot assume that the individual student's ability was precisely measured, nor that the sole factor determining current achievement is past achievement. Students may have disliked a particular subject, and their aversion to it may have caused them to perform poorly on tests used in the classification scheme. It may be incorrect to assume that the aversion remains or is continuing to have an influence on recent measures of achievement.

2) Another problem associated with judging a student's performance in relation to some estimate of ability is known as the regression effect. The regression effect is a statistical artifact that occurs, particularly with individuals at the extreme ends of a score distribution, who tend to regress toward the average due to the less than perfect relationship between ability and achievement measures. For example, individuals classified as A students on an ability measure are statistically likely to be classified at a lower level on a subsequent achievement measure. Conversely, individuals classified as low on the original ability measure are statistically more likely to increase their level on subsequent achievement measures.

2.5 Grading Based on Growth or Improvement.

Previously discussed methods of assigning grades have been based on a pupil's status or level of achievement at the end of a course of study. Grade categories were then decided in either some predetermined, absolute manner, or in terms of the range of achievement or ability within the class. One suggested alternative to such approaches is to assign a grade on the basis of how much improvement a pupil has shown during the learning period. This method appeals to some teachers who feel that the measurement of growth is fairer to student than is the measurement of current status. The appeal is greatest at the early elementary school level when children are learning the basic skills of reading and arithmetic and when pupils are working on individual programs or on a mastery learning approach. However, when improvement is used as the *sole* determinant of grades, serious technical problems arise.

Some Technical Problems with Grading Based on Growth

1) Measuring growth or change requires that a teacher have initial and final scores for each pupil, based ideally, on two different forms of a summative test. If the initial scores are obtained from relatively easy tests, capable students will have limited opportunity to improve on their final assessment. Of course, when pupils know that their grades will be based on the amount of improvement, they soon realize that it is to their advantage to obtain a low score on the initial test (fake ignorance).

In general, developing alternate forms of a test that can reliably measure growth is an extremely difficult task.

2) A problem closely related to the first is that of the regression effect mentioned in an earlier section. Because of the lack of a perfect relationship between the initial and final test results, students who are considerably above or below the average on the initial test tend to regress toward the mean of the distribution on the second test. Consequently, it is much more difficult for students scoring high on the first test to raise their score than it is for students scoring low on the first test.

3) The reliability of gain scores (post-test minus pre-test scores) tend to be very low. Both pre-test and post-test scores have various amounts of random error associated with them, and when the scores are subtracted, the problems related to random error are compounded*.

4) Equal gain scores have different educational meaning at different levels of achievement. This is particularly true when tests are arranged so that the easiest test items precede the more difficult items. A student who improves from 20 percent correct to 30 percent correct on such a test cannot be considered as making the same progress as a second student who improves from 80 percent to 90 percent correct, even though the students have identical gain scores.

At the risk of over-simplifying the problem, compare a student who improves from 15 to 20 words per minute in a junior secondary typing class with one who improved from 40 to 45 words per minute in the same reporting period in the same class. The increase in words per minute is identical, but assigning the students identical grades may be misleading for both students and parents unless the grades are accompanied by comments that relate the achievement of the students to that of the rest of the class.

* For a more complete discussion of this problem, the interested reader is referred to Thorndike and Hagen: "Measurement and Evaluation in Pyschology and Education" Fourth Edition pp 98 to 101.

2.6 Grading Based on Contracts.

Recently, some teachers have begun to organize their courses in such a way that a pupil can, near the beginning of a course, choose and contract for a particular letter grade based on the work he believes he can accomplish. The system is predicated on clearly defined pupil activities, ordered in terms of both quality and quantity of work which increases as the letter grade changes from P to A. Various approaches are used depending upon the nature of the course, the method of instruction,and the characteristics of the pupils in the class.

In general, a course is organized in terms of projects, activities, and tests which are directly related to the learning outcomes of the course and classified in terms of letter grades. High grades are usually awarded to pupils who complete a certain number of required assignments and tests at a designated level of competence plus an optional number of high quality projects. Low grades are awarded to pupils who complete only minimum essentials.

Proponents of the grade contract claim that the procedure allows a teacher to individualize instruction and also reduce competition for grades. Ideally, the contract is based on what one teacher and one student agree in writing will be accomplished during the course for a particular letter grade. On the basis of an explanation by the teacher, and through individual and class discussion, each student decides the grade level to which he aspires, and a contract to this effect is then signed by the student and the teacher at some point in the first few weeks of the course.

Usually the contracts are organized through the use of tasks or points. The former requires the pupil to complete selected tasks plus a certain number of optional tasks to receive a particular letter grade. For example, a contract may read: "for a C, complete the starred activities and any three additional ones; for a B, complete any six additional activities; and for an A, complete any nine additional ones." The point system requires the pupil to accumulate a predetermined number of points to receive a specific letter grade. For example, a pupil might accumulate 100 points for a C, 150 points for a B, 200 points for an A. These point totals are calculated on the basis of the number of points allocated to each of the tasks connected with the course learning outcomes.

NAME_____		DIVISION_____
I will do the following for a letter grade of: (check one)		To demonstrate that I have done these things, I will:
A	1. Develop the following exhibit for the science fair:	1. Display my project at the February 21 science fair according to the rules** Submit weekly progress reports to the instructor in writing.
B	2a. Develop a science notebook on: OR 2b. Visit one of the following science exhibits each month: 1) Planetarium 2) Regional College 3) Local Industry 4) OR 2c. Read one of the following biographies of scientists each month. 1) 2) 3) 4)	2a. Submit notebook to instructor on last day of each month.*** 2b. Submit written report the week after the visit.*** 2c. Submit a book review each month***
C	3. Participate in class activities.	3. Pass oral quiz on laboratory practice*
P	4. Read the textbook. 5. Learn the vocabulary.	4. Pass weekly and six weekly quizzes* 5. Pass weekly spelling and definition quizzes*

*Graded pass or fail - failed exams may be repeated the following day after school.

**See attached sheets.

***Graded pass or fail - failed work may be revised and re-submitted within one week.

SIGNATURE: STUDENT_____ INSTRUCTOR_____

DATE_____

Figure 6. Example of a Letter Grade Contract for a Science Course at the Junior Secondary Level.

In most contract systems, the standards for completing assignments, tasks and tests are spelled out in detail and any unsatisfactory work may be re-done until judged acceptable, provided it is completed within a predetermined time schedule.

If the pupil completes the contract, the agreed grade is assigned. If, however, only part of the agreement is met, only that level of the contract will be binding and the letter grade will be adjusted accordingly. If teacher and student agree, a contract may be revised (for a higher or lower letter grade) at certain stages of the course.

The printed contract, which contains a listing of required tasks or points and their due dates, is prepared in two copies. The teacher and student sign both copies and keep one each for their own use. As example of a letter grade contract used in a science course at the junior secondary level is shown in Figure 6. A student who signed a contract for an A would be required to complete tasks 1 to 5 inclusive. A student who agreed to a contract for a C would complete tasks 3, 4, and 5 only.

The grades used in the grade contract system are basically a method of reporting measurements of achievement in a course. When the learning outcomes of the course are written in specific terms, the measurements can be objective and defined operationally. The grades are also valid. That is, the highest grades go to students who have achieved to the highest degree in terms of the important learning outcomes of the course.

2.7 Grading Based on Criterion-Referencing.

Criterion-referenced grading can be thought of as somewhat equivalent to contract grading. However, with a criterion-referenced approach, the contract signing is omitted. Both methods require the detailed specification of learning outcomes, organized in terms of some hierarchy of complexity, and the assignment of grades based on achievement of tasks and outcomes which have been ordered in a hierarchical fashion.

An example of possible criteria for the assignment of grades in a fictitious unit of a Social Studies course is shown in Figure 7. The "C" worksheet referred to in the figure would cover certain learning outcomes. The "B" and "A" worksheet would differ in quality expected as well as quantity.

Social Studies — Grade 9
European History — 1918-1945
A comparison of the two Wars

Criteria for a Grade of 'C'

C.1. complete the C level worksheets
C.2. complete the C level essay assignment
C.3. obtain an acceptable score on the C level tests*

Above average work on one or more of the criteria will result in a final Grade of C plus.

Failure to complete satisfactorily any one of the criteria will result in a Grade of C minus.

Failure to complete satisfactorily more than one of the criteria will result in a Grade of D.

Criteria for a Grade of 'B'

B.1. complete criteria C.1 and C.2
B.2. complete the 'B' level worksheets**
B.3. complete the 'B' level essay assignment**
B.4. obtain an acceptable score on the B level tests*

Above average work on one or more of B.2 - B.4 will result in a final Grade of B plus.

Failure to complete satisfactorily any one of B.1 - B.4 will result in a Grade of B minus.

Failure to complete more than one of B.1 - B.4 will result in a Grade C or C plus.

Criteria for a Grade of 'A'

A.1. complete criteria C.1 and C.2 and B.2 and B.3
A.2. complete the 'A' level worksheet**
A.3. complete the 'A' level essay assignment**
A.4. obtain an acceptable score on the A level tests*

Above average work on A.2 - A.4 will result in a final Grade of A plus.

**B and A worksheets and assignments would differ from those for C both in quality and quantity expected.

Failure to complete satisfactorily any one of A.1 - A.4 will result in a Grade of A minus.

Failure to complete more than one of A.1 - A.4 will result in a Grade of B or B plus.

Figure 7. Example of Basing Letter Grades on Criterion-Referencing for Social Studies at the Junior Secondary School.

*For suggestions for standard setting see Chapter 5 of "Construction and use of Classroom Tests" (See Bibliography.)

Teachers wishing to use a criterion-referenced grading approach should consider carefully the following comments and suggestions:

1) Organization of the course. The subject matter of the course must be thoroughly understood by the teacher before the course begins. This implies more than just an understanding of the major concepts and principles. It also means understanding the structure of the subject matter and, in particular, of those concepts and principles that the average student finds difficult. Knowing the course structure allows a teacher to better organize it and to establish standards for minimum and other performance levels.

2) Co-operation with others. In schools where multiple sections of a course are taught by different teachers, it would be to teachers' mutual advantage to compare their various standards and decide on a common performance basis for the letter grades that are to be assigned. The possibilities of co-operatively agreeing on criterion-referenced grading standards is one of the main advantages of setting up student tasks in terms of letter grade levels. This process will probably involve teachers working together in the same subject area over a considerable period of time.

3) Investigation of past performance. Performance criteria for different letter grade levels can, in part, be based on the experience teachers have had with previous students. Well organized records of test results, item analyses, individual and class assignments or projects, and the kinds of instructional strategies employed can serve as valuable indicators of what might be expected of groups of students.

4) Distribution of letter grades. Once the criteria for different letter grades are set, a teacher must be prepared to live with the system. Whenever there is no predetermined distribution of letter grades (for example 7% receive A), a teacher must be willing at least theoretically, to assign all A's or all F's. This possibility suggests that it might be wise to conduct a dry run during one semester prior to the implementation of a criterion-referenced grading system in order to rule out the possibility of ending the course with an extraordinary letter grade distribution. If the course is organized so that no pupil ever achieves an A, there is reason to believe that the instructional program is geared to a level below the capability of the classes.

CHAPTER 3

COMBINING EVIDENCE OF ACHIEVEMENT

Some of the previous sections have assumed that letter grades are assigned on the basis of numerical data obtained from various pupil activities such as projects, written reports, class presentations, informal quizzes, and final examinations. Once the numerical data are gathered, two questions are paramount: What relative weight should be given to each component? and How can component scores be combined so that the composite total truthfully reflects the pupil's achievement level?

3.1 Assigning Relative Weight to Components.

The weight allocated to each of the various components used for grading purposes should represent the considered judgement of a teacher or group of teachers. It is here that past experience in teaching the course is of considerable help. An experienced teacher knows the relative difficulty and importance of the various tests and testing situations, the amount of time that pupils have devoted to each task, and will be able to judge each activity in terms of its contribution to further learning in the subject area. All of these professional value judgements will have to be considered when placing a weight on the importance of an activity in relation to its contribution towards a final letter grade.

3.2 Combining Components.

Method 1 using standard deviations. Weighting and combining the various components of achievement are sometimes done very informally, almost at an intuitive level. A common, although not necessarily appropriate procedure is to add all of the scores students have received for tests and assignments that are to be included. For example, using a very simple illustration, suppose a teacher gives only one test and a term project which are to be weighted equally in arriving at letter grades for a class. The procedure using unweighted raw scores is shown for two pupils in Table 5.

TABLE 5

USING UNWEIGHTED RAW SCORES TO DETERMINE A TOTAL

Pupils	Test	Term Project	Total
Ann	50 (1 S.D. above the mean)	50 (the mean)	100
Bob	40 (the mean)	55 (1 S.D. above the mean)	95
Class Data			
mean	40	50	
S.D.	10	5	

The raw score totals suggest that Ann (Total 100) should be considered as achieving at a slightly higher level than Bob (Total 95).

However, a detailed analysis using the class data in the table shows that Ann's higher total was due to the fact that she scored above the mean on the test which had the larger standard deviation. A closer look indicates that Ann and Bob both scored at the class mean on one measure, and at one standard deviation above the mean on the other. If the final totals are to represent the position in the class for both measures weighted equally, then both students should have the same totals.

In general, calculating a total by adding the raw scores of different measures does not necessarily weight the different measures equally. When scores are combined in this way, those with the greater variability (standard deviation) have greater weight.

A practical way to determine the actual weight each measure contributes towards the total is to consider each measure's contributions as approximately equal to the size of its standard deviation*. If a teacher wishes to ensure that each test or project contributes an equal amount to the variability of the composite total, an adjustment would have to be made to the pupils' raw scores. Various procedures are available. Precise methods would involve converting each raw score into some type of standard score so that each measure would have the same variability.

In this case, the simplest method to ensure that the test and the term project contribute equally is to multiply the term project scores by 2 as shown in Table 6. This is done because according to the standard deviations in Table 5, the Test scores contribute twice as much to the total score variability as the Term Project.

TABLE 6

USING WEIGHTED RAW SCORES TO DETERMINE A TOTAL

Pupils	Test	Term Project (weighted data)	Total
Ann	50	2 x 50 = 100	150
Bob	40	2 x 55 = 110	150
Class Data			
Mean	40	100	
S.D.	10	10	

(Based on Table 5)

*For a straightforward method of calculating the standard deviation, see Chapter 4 of Construction and Use of Classroom Tests: A Resource Book For Teachers (distributed by the Learning Assessment Branch).

The totals are now equal and reflect each pupil's average position in the class based upon the two tests being given equal weight.

Method 2 Using Stanines. In many practical situations, it is convenient to convert all raw scores from various measures into stanines*. When this is done, because each stanine distribution has the same variability, the stanines can be weighted and a composite total may be obtained. The formula for calculating the composite total (CT) is:

$$CT = \Sigma WS$$

where CT = composite total
Σ = the sum
W = weight for each component measure
S = a pupil's stanine for each measure

Consider the following example for a student whose final grade is based on combining five measurements expressed in stanines and weighted as shown in column two.

Measure	Weight (W)	Stanine (S)	WS
class reports	1	6	6
daily work	1	6	6
project	2	7	14
mid-term test	2	5	10
final exam	3	6	18

Composite Total CT = (1 x 6) + (1 x 6) + (2 x 7) + (2 x 5) + (3 x 6)
= 6 + 6 + 14 + 10 + 18
= 54

The composite total calculated in this way for each student can be compared to the totals of other students in the class to provide a ranking of achievement for the class. Assigning the letter grades from such a ranking, (using the techniques described in Chapter 2) reflects the desired weighting of the scores for the several activities and measures completed throughout the course.

It should be noted that in addition to providing a convenient method of eventually combining scores, stanines may also be useful for interpreting norm-referenced classroom tests.

*Procedures for calculating stanines are presented in the Appendix.

CHAPTER 4

GRADE COMPARABILITY

One of the most frequent complaints about grades is that they are not comparable across teachers or subject areas. One teacher might, for example, give 14 A's to an English 8 class while another teacher of the same subject may assign only 3 A's, and yet the classes of 30 pupils may be very similar in scholastic ability and presumably are expected to achieve the same learning outcomes. Such divergent grade distributions can produce a considerable amount of antagonism within a school. If such differences persist from year to year, they can destroy parents' and others' confidence in the meaning of particular letter grades. What can be done, then, within a school or district, to help reduce the variability of grades across courses or within different sections of a particular course? The following suggestions are offered for consideration:

1. Publish and discuss letter grade distributions.

Teachers in a school or school district can profit from periodic discussions of grading standards. By spending time sharing and considering their varying philosophical orientations, teachers can move towards a common policy. Greater similarity between teachers in assigning grades should result if discussions are based on published grade distributions.

Staff meetings might include discussions of published letter grade totals organized by subject areas (e.g. Commerce, Physical Education, Mathematics, Fine Arts etc.). Subject department meetings might consider letter grade totals organized by course (e.g. Math. 8, Math. 9, Math. 10, Algebra 11 etc.). If all teachers in a department are willing, letter grades organized by teachers may be a topic of a productive meeting. Published lists of letter grades organized by grade level or by subject area of secondary schools might be considered as an agenda item for a district-wide meeting of administrators.

Teachers of the same subject or grade could discuss grading standards after they have graded the same set of student tests or projects. The sample of student work should be chosen in advance to represent all expected levels of achievement in that subject or grade. The marks assigned to these tests or projects by the teachers would serve as the catalyst for discussions of the similarities and differences of standards and expectations.

2. Use of an anchor test.

Some schools have attempted to develop procedures that increase the comparability of grades across teachers or across classes of the same course through the use of some "anchor". That is, the letter grade assignment for a class is anchored to a variable independent of the specific achievement in the class during the reporting period. The

"anchor" might be the previous letter grade average or a distribution of scholastic ability test results within the class. Such procedures usually help to increase the comparability and meaningfulness of grades for courses with a common curriculum and with large heterogeneous pupil groups. However, they run into difficulties when the classes are small or specialized and the cautions in section 2.4 of this booklet apply to this procedure.

3. Graph internal and external grades.

Teachers wishing to compare the grades they assign to their students with those assigned by the other teachers who have the same students, should consider using a helpful graphical technique, the histogram. The procedure is best described through an example.

Table 7 shows the English marks assigned by a grade 11 teacher (internal grades) and those assigned to the same ten students by the teachers of Mathematics, Chemistry, Social Studies, French, Physics, Biology, (external grades) in the same reporting period.

TABLE 7

INTERNAL AND EXTERNAL GRADES FOR TEN STUDENTS

	Internal Grade (English)	External Grade		Internal Grade (English)	External Grade
Student 1	A 1.00	Math. B .25 Chem. C .25 Soc. C .25 Fren. C .25	Student 6	B 1.00	Chem. C .33 Soc. B .33 Fren. P .33
Student 2	A 1.00	Math. C .33 Chem. B .33 Soc. C .33	Student 7	B 1.00	Fren. C .25 Chem. F .25 Math. F .25 Soc. A .25
Student 3	A 1.00	Phys. B .25 Chem. C .25 Soc. B .25 Biol. C .25	Student 8	C 1.00	Biol. B .33 Math. P .33 Fren. P .33
Student 4	A 1.00	Phys. C .50 Chem. C .50	Student 9	C 1.00	Math. A .50 Soc. P .50
Student 5	B 1.00	Math. F .25 Soc. C .25 Chem. B .25 Biol. A .25	Student 10	P 1.00	Soc. F .25 Fren. P .25 Math. C .25 Phys. P .25

Step 1: Count the number of internal A's, B's, C's and P's assigned (in Table 7 there are 4, 3, 2 and 1 respectively).

Step 2: Draw a histogram to represent this grade distribution as shown on the left of Figure 8 below. Each internal A grade has contributed one unit of area to the A column of the graph. Similarly each internal B, C and P grade has contributed one unit to the histogram.

Step 3: Consider the external A grades assigned. There are three such grades in Table 7, but the A's of Students 5 and 7 are weighted as .25 because they each take four subjects other than English, whereas the A assigned to Student 9 accounts for .50 because this student only has two other subjects. (Thus, regardless of the number of courses taken, each student contributes one unit to the English graph and one unit in total to the graph of the external grades). The total weighting for external A's is therefore .25 +.25 +.50 = 1.0.

Step 4: Repeat Step 3 for external grades B, C, P, and F. From Table 7 the weightings are as follows:

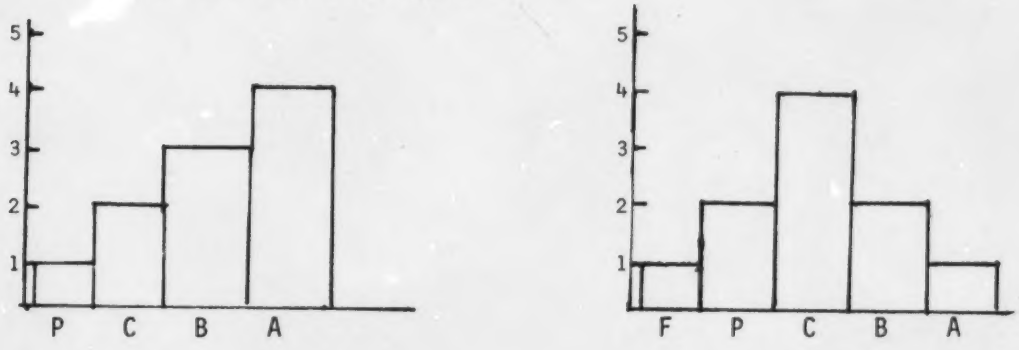
B is .25 + .33 + .25 + .25 + .25 + .33 + .33 = 2.0

C is .25 + .25 + .25 + .33 + .33 + .25 + .25 + .50 + .50 + .25 + .33 + .25 + .25 = 4.0

P is .33 + .33 + .33 + .50 + .25 + .25 = 2.0

F is .25 + .25 + .25 + .25 + 1.0

Step 5: Draw the histogram to represent the external grades as shown on the right of Figure 8.



Internal Grade Distribution External Grade Distribution

Figure 8. Histograms Showing Internal and External Grade Distributions

A brief look at the histograms reveals differences between the two distributions. Whether the English grades are too lenient is a question that must be solved by further investigation. However, the histogram approach does provide a teacher with a useful technique for revealing relevant differences in grading standards.

CHAPTER 5

DISCUSSION QUESTIONS

The questions which follow, based primarily on the preceding chapters, are intended to stimulate fruitful examination of the grading practices in a school or school district. There are no "right" answers to many of the questions, but the discussions generated by them may focus the attention of the appropriate personnel on existing problems and on potential solutions.

I Issues of Particular Concern within a School

(a) The basis for grading:

* Is the grading system based upon a clear statement of the important learning outcomes of each course or subject?

* Is the grading system designed to report final achievement status in terms of these important learning outcomes?

* Do students know in advance of the basis used for assigning letter grades?

* Do teachers use a wide range of relevant and appropriate sources of achievement indicators as a basis for assigning letter grades?

* Are teachers aware of the advantages and disadvantages of assigning letter grades on the basis of:

    a) the normal distribution?
    b) natural breaks?
    c) growth or improvement?
    d) written contracts?
    e) criterion-referencing?

* Are teachers aware of a variety of techniques for weighting and combining evidence of student achievement when assigning letter grades?

(b) Grade Comparability:

* Is there agreement amongst teachers of the same subject or grade as to what should be taken into account in assigning letter grades?

* Is there agreement amongst teachers as to what factors should be taken into account in determining the distribution of letter grades for different levels of instruction and for different subject area?

* Are variations from class to class in both ability and achievement taken into consideration when assigning grades?

* Are systematic procedures planned to study the causes of, and the extent of, variability in grading standards and letter grade distribution from teacher to teacher, and/or from school to school?

II Issues of Broad General Concern

* Is the grading and reporting system simple and understandable?

* Does the grading system make optimum contribution to the students and to student learning?

* Does the grading and reporting system provide the basis for making fruitful decisions about student progress?

* Is the reporting format capable of communicating the meaning of the letter grades to the potential users?

* Are steps regularly taken to ensure that all students, parents and employers understand the grading system and the reporting format?

* Is there an organized committee of administrators, teachers, parents, students and school board members co-operatively studying the grading and reporting practices and making recommendations for improvement?

* Are unique and alternative grading and reporting systems available for exceptional or atypical students?

* Are non-achievement characteristics such as attitude, behaviour, and attendance separated from the achievement category in the reporting format?

* Have alternatives to the conventional letter grading system been considered for courses that use non-quantified factors (such as attitude and effort) in reporting student progress?

APPENDIX

CONVERTING A SET OF SCORES INTO STANINES

The stanine (short for standard-nine) scale is a system of standard scores that divides a raw score distribution into nine categories. Only integers 1 through 9 are used with stanine 1 referring to the lowest group of raw scores, stanine 5 in the middle group and stanine 9 the highest group. Each stanine, except 1 and 9, includes a band of raw scores one-half a standard deviation wide when referenced to the normal curve.

Changing a set of raw scores into stanines forces the original distribution into a normal distribution with a mean of 5 and a standard deviation of 2. Technically, the procedure is known as an area transformation (in contrast to a linear transformation). The procedure is facilitated by the use of the Stanine Table, on the following page, which lists the number of scores that should be included in each stanine category for a given number of students. The table also contains the percentage of students that fall within each stanine as well as the cumulative percentages.

Stanine Table Showing Number of Scores Falling in Each Stanine for Various Class Sizes

	STANINES								
	1	2	3	4	5	6	7	8	9
Percent of Scores	4%	7%	12%	17%	20%	17%	12%	7%	4%
Approximate Range of Percentile Ranks	Below 5	5-11	12-23	24-40	41-60	61-77	78-89	90-96	Above 96
Size of Class									
20	1	1	2	4	4	4	2	1	1
21	1	1	2	4	5	4	2	1	1
22	1	2	2	4	4	4	2	2	1
23	1	2	2	4	5	4	2	2	1
24	1	2	3	4	4	4	3	2	1
25	1	2	3	4	5	4	3	2	1
26	1	2	3	4	6	4	3	2	1
27	1	2	3	5	5	5	3	2	1
28	1	2	3	5	6	5	3	2	1
29	1	2	4	5	5	5	4	2	1
30	1	2	4	5	6	5	4	2	1
31	1	2	4	5	7	5	4	2	1
32	1	2	4	6	6	6	4	2	1
33	1	2	4	6	7	6	4	2	1
34	1	3	4	6	6	6	4	3	1
35	1	3	4	6	7	6	4	3	1
36	1	3	4	6	8	6	4	3	1
37	2	3	4	6	7	6	4	3	2
38	1	3	5	6	8	6	5	3	1
39	1	3	5	7	7	7	5	3	1
40	1	3	5	7	8	7	5	3	1
50	2	3	6	9	10	9	6	3	2
60	3	4	7	10	12	10	7	4	3
70	3	5	8	12	14	12	8	5	3
80	3	6	9	14	16	14	9	6	3
90	4	6	11	15	18	15	11	6	4
100	4	7	12	17	20	17	12	7	4

The following distribution of thirty scores is used to demonstrate the procedure for converting a set of scores into stanines.

Score	f	Score	f
40	1	30	3
39	2	29	1
37	2	27	2
35	1	26	2
34	2	25	5
33	2	23	1
32	3	21	1
31	1	19	1

The first step is to set up a list of the frequencies desired for stanines, 9, 8, and so on down. These frequencies can be read from the row opposite the class size of 30 in the Stanine Table. They are listed in the theoretical frequency column below.

Stanine	Theoretical Frequency	Actual Frequency	Equivalent Raw Scores
9	1	1	40
8	2	2	39,39
7	4	3	35,37,37
6	5	7	32,32,32,33,33,34,34
5	6	5	29,30,30,30,31
4	5	4	26,26,27,27
3	4	5	25,25,25,25,25
2	2	2	21,23
1	1	1	19

From the theoretical frequency we see that we need to include just one raw score for both stanines 1 and 9. Therefore, we assign stanine 1 to the lowest score, 19. We also assign stanine 9 to the highest score, 40. We then determine the raw scores that should be included in stanines 8 and 2 (working down from the top and up from the bottom). We need two scores for both stanines 2 and 8. We note that score 39 has a frequency of two and therefore we assign stanine 8 to score 39. Similarly since score 21 and 23 each have a frequency of one, we assign stanine 2 to each of these scores. Other stanines are assigned in a similar manner.

Sometimes one is confronted with the problem of whether or not to assign to a stanine category a particular raw score that is near the cut-off point between stanine categories and which will result in the stanine category having either a larger or smaller frequency than the required theoretical frequency. The simplest way to solve the problem is to calculate the percentile rank of the score and then assign the stanine category that includes the particular percentile rank as shown in the third row of the stanine table.

In general (although not in the example above) working from both extremes of the distribution will sometimes cause stanine 5 to have an actual frequency that deviates the most from the theoretical frequency. Because it has the highest theoretical frequency, stanine 5 can tolerate a larger discrepency between the actual and theoretical frequencies than the other categories. The actual frequency at each stanine level should be as close as possible to the theoretical values. Remember, however, that all equal scores MUST be assigned the same stanine. That is, everyone having the same raw score must have the same stanine.

In general, if the raw scores are not well spread out and, as well, are non-symmetrical around the mean, as is the case in the example above, the actual frequencies will deviate from the theoritcal frequencies.

GLOSSARY

achievement test A test that measures the extent to which a person has "achieved" something acquired certain information, or mastered certain skills — usually as a result of planned instruction or training.

arithmetic mean A kind of average usually referred to as the mean. It is obtained by dividing the sum of a set of scores by their number.

average A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the arithmetic mean.

distribution (frequency distribution) A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

grade inflation A trend wherein the average level of grades assigned within a course, department, or institution increases over time.

item A single question or exercise in a test.

item analysis The process of evaluating single test items in respect to certain characteristics. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some external criterion.

mean (M) See arithmetic mean.

median (Md) The middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scores are below the median and half above it, except when the median itself is one of the obtained scores.

negatively skewed distribution A distribution in which there is a greater range of scores below the mean than above it; apparent from visual inspection of a frequency curve in that it has a long tail extending to the low scores without a corresponding tail extending to the high scores.

normal distribution A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. In such a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. Mean and median are identical. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

norms Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

percentile (P) A point (score) in a distribution at or below the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile (P_{35}) is regarded as equalling or surpassing 35 percent of the persons in the group. It also means that 65 percent of the performances exceed this score. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

percentile rank (PR) The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

positively skewed distribution A distribution in which there is a greater range of scores above the mean than below it; apparent upon visual inspection of a frequency curve in that it has a long tail extending to the high scores without a corresponding tail extending to the low scores.

range For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or variability, since it is based upon only two extreme scores. Range is also used in reference to the possible spread of measurement a test provides, which in most instances is the number of items in the test.

raw score The first quantitative result obtained in a scoring test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

reliability The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of reliability coefficient or by the standard error or measurement derived from it.

skewed distribution A distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

standard deviation (S.D.) A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. is sometimes called "sigma" and is represented by the symbol

standard score A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a z-score, is an expression of the deviation of a score from the mean score of the group in relation to the standard deviation of the scores of the group. Thus:

$$\text{standard score (Z)} = \frac{\text{raw score (X) - mean (M)}}{\text{standard deviation (S.D.)}}$$

Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty; also, successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible.

standardized test (standard test) A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term "standardized" to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others would add "commercially published" and/or for "general use".

stanine One of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale has values from 1 to 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is ½ S.D. in width, with the middle (average) stanine of 5 extending from ¼ S.D. below to ¼ S.D. above the mean.

summative evaluation Summative evaluation in the classroom generally is used to refer to evaluative procedures (tests, examinations, reports, projects) conducted customarily at the end of major units for the purpose of assessing a student's performance in relation to others and/or to a predetermined criterion level. These results normally form a substantial basis for the student's course grade. An additional purpose of summative evaluation is to provide the teacher with information concerning the relative effectiveness of the preceding unit in meeting the designated instructional objectives.

validity The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional "extent to which a test measures what it is supposed to measure," since the validity of a test is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(1) content, curricular validity For achievement tests, validity is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of educational goals; and often by analysis of the processes required in making correct responses to the items. Face validity, referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) criterion-related validity The extend to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some given criterion measure. Predictive validity refers to area, as evidenced by correlations between scores on the test and future cirterion measures of success (e.g., the relation of score on an academic aptitude test administered in high school to grade point average over four years of college). In concurrent validity, no significant time interval elapses between administration of the test to one generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score would be.

(3) construct validity The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluation the validity of tests that have been constructed on the basis of an analysis (often factor analysis) of the nature of the trait and its manifestations. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

variability The spread or dispersion of test scores, best indicated by their standard deviation.

BIBLIOGRAPHY

Ebel, R. L. Essentials of Educational Measurement. 3rd ed. Prentice-Hall, 1979. See Chapter 12 "Marks and Marking Systems".

Gronlund, N.E. Improving Marking and Reporting in Classroom Instruction. Macmillan, 1974.

Gronlund, N.E. Measurement and Evaluation in Teaching. 3rd ed. Macmillan, 1976. See Chapter 19 "Marking and Reporting".

Hills, J.R. Measurement and Evaluation in the Classroom. Charles E. Merrill, 1976. See Part 3 "Grading and Marking".

Kirschenbaum, M., R. Napier and S. Simon. Wad-ja-get? Hart, 1971.

Mehrens, W.A. and I.J. Lehmann. Measurement and Evaluation in Education and Psychology. Holt, Rinehart and Winston, 1978. See Chapter 17 "Marking and Reporting the Results of Measurement".

Simon, S.B. and J.A. Bellanca, editors. Degrading the Grading Myths: A Primer to Alternatives to Grades and Marks. Association for Supervision and Curriculum Development, 1976.

Terwilliger, J.S. Assigning Grades to Students. Scott, Foresman, 1971.

Thorndike, R.L. and E. Hagen. Measurement and Evaluation in Psychology and Education. 4th ed. John Wiley and Sons, 1977. See Chapter 16 "Marks and Marking".

Relevant Publications of the Learning Assessment Branch,

Ministry of Education, Science and Technology

Construction and Uses of Classroom Tests: A Resource Book For Teachers

Selection and Use of Standardized Tests: A Resource Book for School and District Personnel.